Joint linear modeling of transcriptomics and proteomics is predictive of cancer metastasis

Ronit Sharma*, Nikolaos Meimetis*, Arjana Begzati, Shashwat Depali Nagar, Benjamin Kellman, Hratch M. Baghdassarian
*contributed equally

Author Information

Author Affiliations

Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
Nikolaos Meimetis, Hratch M. Baghdassarian

Bioinformatics and Systems Biology Graduate Program, University of California, San Diego, La Jolla, CA, 92093, USA
Arjana Begzati

School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, 30332 USA
Shashwat Depali Nagar

Ragon Institute of MGH, MIT, and Harvard, Cambridge, MA, 02139 USA
Benjamin Kellman

Unaffiliated
Ronit Sharma

Contributions

R.S.: Conceptualization, Transcriptomic Analysis, Writing, and Editing N.M.: Conceptualization, Neural Network Analysis, Writing, and Editing A.B.: Proteomic and Feature Analysis, Writing S.D.N.: Power Analysis, Writing B.K.: Conceptualization H.M.B.: Conceptualization, Analysis, Writing, Editing, and Supervision.

Corresponding authors

Correspondence to Hratch M. Baghdassarian (hmbaghda@mit.edu).

Abstract

Cancer metastasis, a process in which cancer cells migrate to secondary sites, accounts for 90% of cancer deaths. While machine learning has been used to predict metastasis in a variety of ways, they tend to be specific to tumor types or classification tasks. Here, we provide a pan-cancer machine learning model that is easily interpretable and directly links the transcriptomic and proteomic profiles of hundreds of cancer cell lines from DepMap with metastasis in a tissue-agnostic manner. We show that, for the relatively small sample size here, linear models perform just as well as non-linear models, including neural networks. We also demonstrate that transcriptomics outperforms proteomics for predictive tasks. However, the combination of both transcriptomics and proteomics leads to an overall increase in performance, indicating that these modalities can complement each other. Finally, we demonstrate that those features identified as most important by the joint omics linear model fit are over-represented in biological pathways related to metastasis, suggesting that the model learns relevant information that may be useful in identifying universal biomarkers or therapeutic targets of cancer metastasis.

Introduction

Cancer is the second-leading cause of death in the United States, and metastasis accounts for 90% of these deaths. During metastasis, cancer cells with increased motility detach from the primary tumor, migrate, and invade and adhere to organs. To enable this, metastatic cells undergo the epithelial–mesenchymal transition and develop specific properties such as resistance against anoikis and the formation of invadopodia. Elucidating the molecular mechanisms underlying these various properties favoring metastasis has motivated multiple therapeutic approaches to treat cancer[1–3]. As such, identifying molecular features in tumors predictive of metastasis in a data-driven manner can reveal novel genes associated with metastatic properties and can have implications for therapeutic targets.

Cell motility, a fundamental cellular behavior of metastasis, varies substantially across the tumor microenvironment and cell lines[4]. Metastasis changes according to contexts such as genetics, extracellular cues, and spatial location that alter the molecular state of cancer cells[5,6]. Consequently, the combinatorial space of molecular features underlying metastatic phenotypes is difficult to decipher[7], and it is difficult to identify generalizable mechanisms of metastasis.

Systems biology enables the study of such complex functions, treating the cell as a system that receives input signals, processes them according to their state, and outputs a response. Thus, computational models can help elucidate how molecular and phenotypic profiles are related[8–10]. In particular, machine learning (ML) approaches can leverage large omics datasets[11] to train predictive models that, for example, identify disease biomarkers, provide insights to gene regulation, and guide the designing of drugs[12].

Such approaches have been used specifically to study cancer metastasis. For example, migration rates have been predicted from feature sets that do not involve global molecular profiles, such as cell morphology[13]. Additionally, there is extensive work on predicting metastasis from omics for tumors in specific tissues or cancer types[14]. While some pan-cancer approaches exist, they are focused on classification tasks, and some also use complex neural nets that are not easily interpretable[15,16].

Here, instead, we provide a prediction for the likelihood of cancer metastasis as a continuous variable using an easily interpretable linear model. We leverage publically available transcriptomic, proteomic, and phenotype data from DepMap[17,18] and MetMap[19], which spans 21 types of solid tumor, indicating that our approach is generalizable and, since we do not incorporate the tissue-specific information reported in MetMap, tissue-agnostic. Our analysis of transcriptomics measurements shows that, on these data, simple linear models can perform as well as more complex, state-of-the-art non-linear ML approaches. Additionally, while transcriptomics seems to be the most informative modality, we observed that the combination of both transcriptomics and proteomics may increase predictive performance. Finally, we identify that specific features from these modalities are combined by the model in a complementary manner, with priority going to those associated with biological pathways relevant to metastasis.

Results

In order to determine whether the global abundance of gene products is informative of metastasis, we leveraged DepMap's cell line measurements[17]. Specifically, we fit a dataset containing 481 samples on a set of machine learning models that take as input protein-coding genes measured by RNA-sequencing and predict as output metastatic potential reported by MetMap 500[19]. For each model, we conducted hyperparameter tuning with Optuna[20] using nested cross-fold validation (CV) (see Methods for details). Despite potentially complex relationships between molecular features and cell phenotypes such as cancer cell metastasis, we observed no significant differences between non-linear and linear models (Mann-Whitney U, Benjamini-Hochberg FDR correction, $q \geq 0.86$). (Fig. 1a, Table 1).

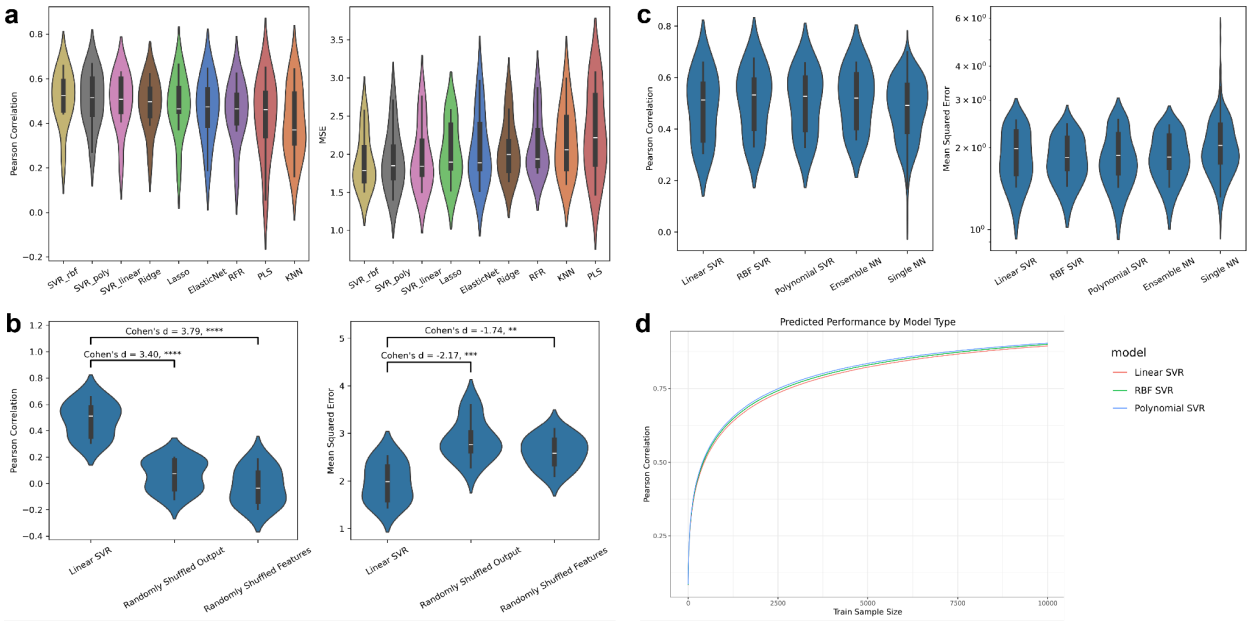

Fig. 1: Linear models exhibit predictive power for metastasis from transcriptomics. All statistical significance is determined using a Mann-Whitney U (MWU) test and the FDR is controlled for multiple testing using the Benjamini-Hochberg correction (**** $q \leq 10^{-4}$, *** $q \leq 10^{-3}$, ** $q \leq 0.01$, * $q \leq 0.1$). Comparisons that are not significant are not annotated. Statistical tests are two-sided unless otherwise specified. Models' predictive performance is assessed either by Pearson correlation (left panels) or mean squared error (right panels). Violin plots visualize performance distribution output from 10-fold CV. **a** Performance of 9 predictive models (x-axis) is ordered by best mean value. For each fold, specific model hyperparameters were optimized using nested 5-fold CV. No comparisons have FDR ≤ 0.1. **b** Consensus Linear SVR model performance compared to random baseline models (x-axis) using 10-fold CV (one-sided MWU test). **c** Comparison of consensus linear SVR model to other consensus SVR models and neural nets (x-axis) performance using 10-fold CV. No comparisons have FDR ≤ 0.1. **d** Estimates of Pearson correlation (y-axis) for each consensus model type across training sample size (x-axis) from 0 to 10,000 training samples. Estimates are derived from a power function regression modeling Fisher z-transformed Pearson correlation as a function of Train Sample Size and including model type as a covariate.

Since support vector regression (SVR) models were the top performers, and because non-linear kernels did not outperform the more interpretable linear kernel, we proceeded with downstream analyses using the linear SVR. Upon identifying a consensus model (see Methods for details), we observed a mean Pearson correlation of 0.48 and mean MSE of 1.97, outperforming random baselines in which the output values or input features were randomly permuted (Fig. 1b). To further confirm our observation that nonlinear models do not outperform linear models, we compared our consensus linear SVR performance with that of consensus top-performing SVRs that use a radial basis function kernel and a polynomial kernel. Additionally, we compared it to a single and an ensemble of ANNs. Again, we found no significant differences in performance between these non-linear models and the linear SVR (Fig. 1c).

Finally, because the number of features used is much larger than the number of available samples, we conducted a power analysis to determine how each models' performance changes

with sample size (Fig. S1, see Methods for details). First, we determined that the change in performance across sample size is best explained by a power function, as opposed to a linear or exponential function (Table 2, see Methods for Details). Next, we fit a power function regressing model performance against training sample size, and we included the model type as a covariate to assess differences between them (Fig. S2). When controlling for sample size, this regression estimates the nonlinear SVRs have a slightly ($10^{0.01}$ = 1.02-fold) higher Fisher z-transformed Pearson correlation than the linear SVR. However, practically speaking, this difference is negligible (Fig. 1d). For example, when extrapolating to 10,000 samples, all model types are estimated to have ~0.90 Pearson correlation. Results for MSE are qualitatively similar (Fig. S2). This indicates that the sample size is a much more substantial limiting factor in model performance than model type, and that linear models may continue to perform as well as non-linear models even at larger sample sizes.

Altogether, these results indicate that machine learning models can use transcriptomics to explain ~23.5% of the variance in metastatic potential, and that, for these data, more interpretable linear models suffice in capturing this relationship.

Next, because protein-level gene abundance may contain additional information about cell phenotypes independent of that available in RNA-level gene abundance[21], we asked how our models may perform on proteomics data. Following the same process as with transcriptomics (see Methods for details), we identified a set of top-performing consensus models for proteomics data (Ridge and support vector machines) (Fig. S3a). Because the proteomics data only had 248 samples, we compared it to the transcriptomics consensus linear SVR model run on an equal sample size from the power analysis. Using 10-fold CV, we found that the transcriptomics model outperformed the proteomics models as measured by MSE (Fig. 2a). Additionally, proteomics models did not outperform the transcriptomics one or performed worse than it as measured by Pearson correlation. The best consensus models for proteomics were Ridge and SVR with polynomial kernels for Pearson correlation and MSE, respectively, with mean values of 0.40 and 2.59. In comparison, the consensus linear SVR for transcriptomics at the same sample size had a Pearson correlation and MSE of 0.44 and 2.21, respectively. Finally, unlike the transcriptomics model, some of the consensus proteomics models did not perform better than random baselines as measured by MSE (Fig. S4); interestingly, these were the best performing Ridge and SVR with a polynomial kernel. Altogether, this indicates that transcriptomics may be more informative of cancer cell metastasis as compared to proteomics.

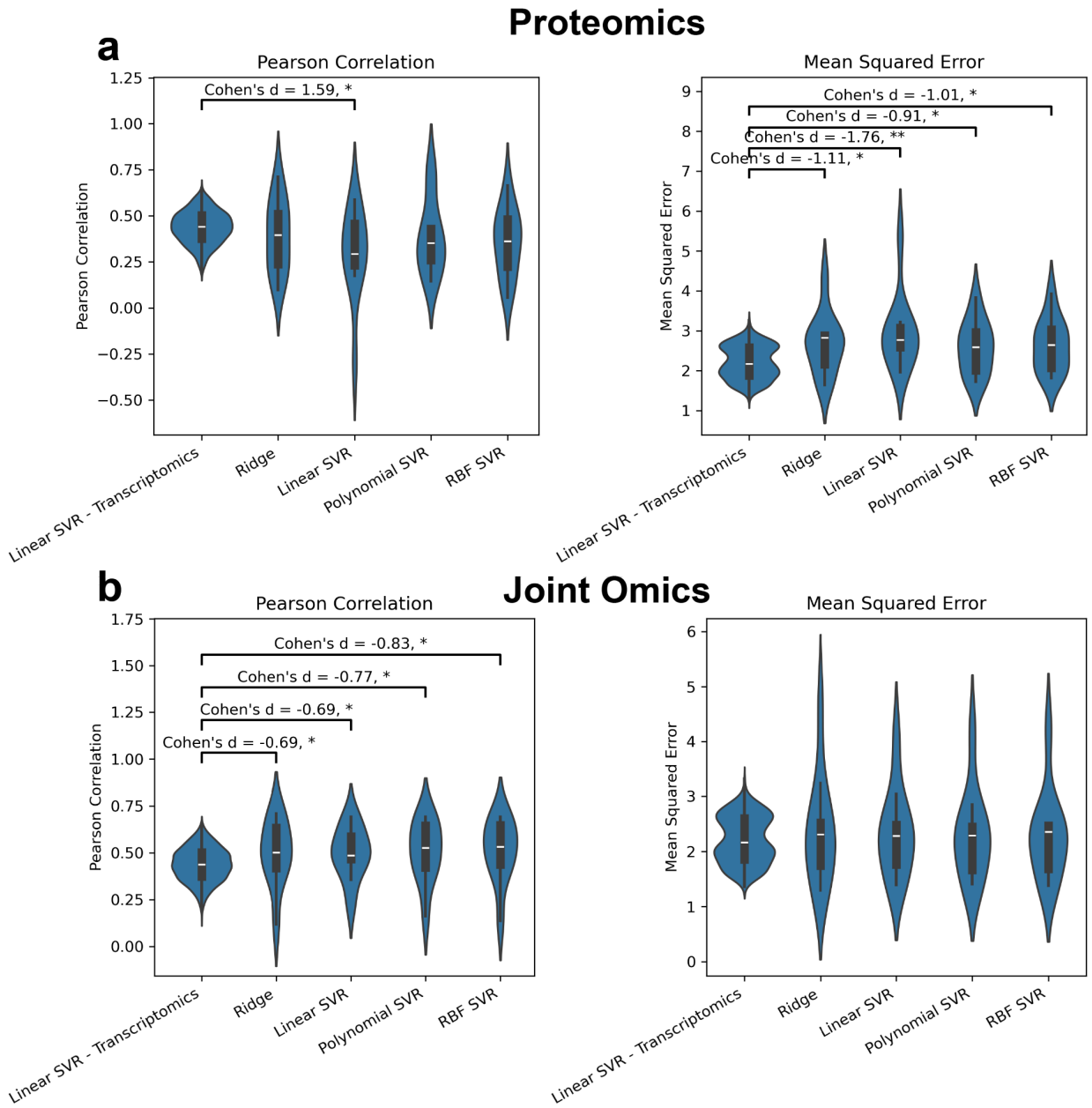

Fig. 2: Multi-omic modeling of cancer cell metastasis demonstrates discrepancies between transcriptomics and proteomics. All statistical significance is determined using a Mann-Whitney U (MWU) test and the FDR is controlled for multiple testing using the Benjamini-Hochberg correction (**** $q \leq 10^{-4}$, *** $q \leq 10^{-3}$, ** $q \leq 0.01$, * $q \leq 0.1$). Comparisons that are not significant are not annotated. Statistical tests are one-sided (comparing the transcriptomics only model to other modalities). Models' predictive performance is assessed either by Pearson correlation (left panels) or mean squared error (right panels). Violin plots visualize performance distribution output from 10-fold CV. Violin Plots of the performance of consensus linear SVR model at the same sample size as **a** proteomics data (248 samples) or **b** joint omics data (247 samples) assessed by power analysis and top-performing consensus proteomics models.

We then asked whether incorporating both transcriptomics and proteomics could enhance models' predictive performance. After again identifying top-performing consensus models (Fig. S3b, see Methods for details), we found that the joint omics models outperformed the consensus linear SVR transcriptomics model as measured by Pearson correlation (Fig. 2b). The best consensus model for joint omics (SVR with radial basis function kernel) had a mean Pearson correlation and MSE of 0.51 and 2.28, respectively, as compared to the consensus linear SVR for transcriptomics at the same sample size having a Pearson correlation and MSE of 0.44 and 2.21, respectively. All consensus joint omics models performed better than random baselines (Fig. S5). However, effect sizes here were small (-0.69 ≥ Cohen's d ≥ -0.83 for Pearson correlation), joint omics models did not outperform transcriptomics as measured by MSE, and the transcriptomics model had lower standard deviations (0.09 Pearson correlation, 0.43 MSE) than consensus joint omics models (minimum across consensus models of 0.14 Pearson correlation, 0.79 MSE).

Nevertheless, since joint omics offered a slight boost in performance, we asked how these molecular features were combined by the model to inform cancer cell metastasis. We proceeded in this analysis with the consensus linear SVR as it had the lowest standard deviation in performance for both metrics during 10-fold CV; additionally, it was the top-performing linear model and the other top-performing, less interpretable non-linear models did not outperform the linear SVR (Fig. S6). We rank-ordered features by the absolute value of their coefficient in the model fitted to the full dataset. We first ran an over-representation analysis (ORA) of the top 500 features, separated by those with negative or positive coefficients (Fig. S7). Many of the overrepresented gene sets are related to metastasis. For example, positive regulation of cell migration is any process that activates or increases cell migration, a key feature of metastatic cancer cells. Additionally, chemotactic migration (chemotaxis) is required by cancer cells to invade the surrounding tissues of the tumor, with EGF and other growth factors serving as factors driving cancer cells to respond to this stimulus and migrate (cellular response to growth factor stimulus)[22]. Integrins play a key role in cellular adhesion, as well as intracellular signaling pathways that regulate cytoskeletal organization and cellular death, all of which are strongly associated with cancer metastasis[23]. Finally, Reactive Oxygen Species (ROS) metabolism can trigger signaling pathways that lead to cellular migration[24].

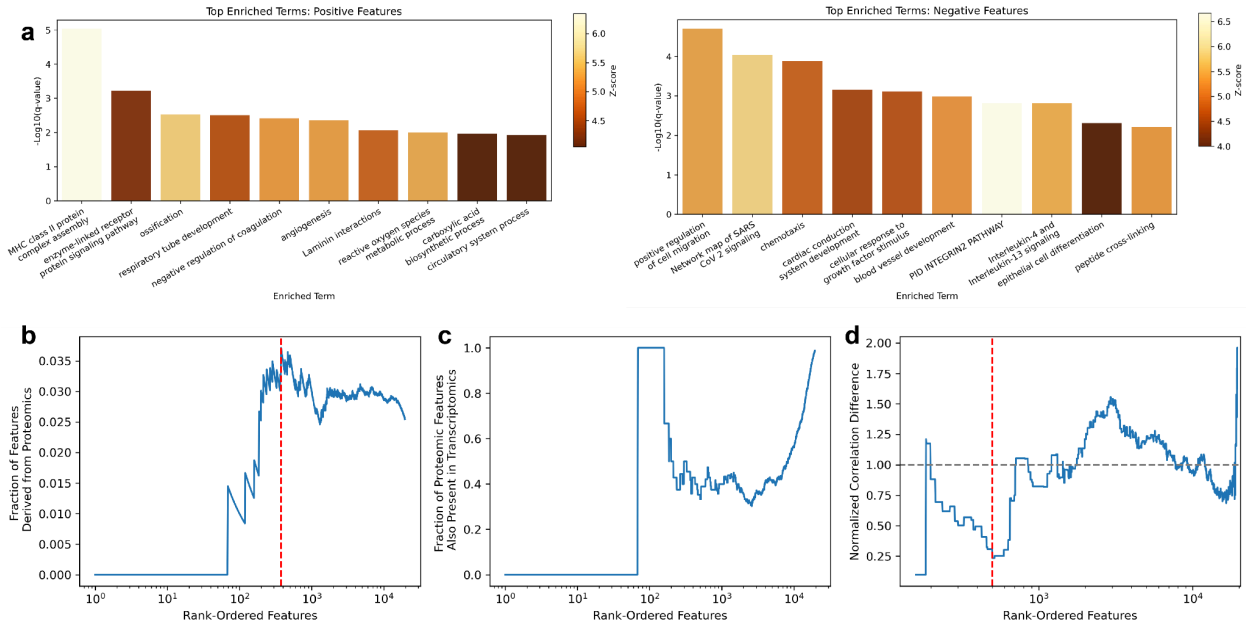

Fig 3. Multi-omic modeling captures coordination between proteomic and transcriptomic features associated with metastasis. a Barplots of the top 10 Metascape over-represented summary terms for features with negative (left panel) and positive (right panel) coefficients in the consensus linear SVR. Terms are rank-ordered by $\log_{10}$(q-value) (y-axis) and colored by z-score. Z-scores are calculated as the average across individual members of each summary term. **b** Line Plots of the fraction of total features derived from proteomics (y-axis) as a function of the number of features at a given rank-order (x-axis). Features are rank-ordered by absolute value of the model coefficient. Red-dashed line represents the rank at which the proteomic feature fraction is maximized. **c** Line Plots of the fraction of total proteomic features (y-axis) that overlap with transcriptomics as a function of the number of features at a given rank-order (x-axis). For a proteomic feature at a given feature rank, we define a corresponding overlapping transcriptomic feature as one that has a shared gene product and a feature rank at or below ("higher rank", more important) that of the proteomic feature. Conversely, a non-overlapping transcriptomic feature is one that has a shared gene product with the proteomic feature and a feature rank above ("lower rank", less important) that of the proteomic feature. **d** Line Plots comparing differences in proteomic-to-transcriptomic feature correlation for overlapping or non-overlapping genes (y-axis) a given rank-order (x-axis). The correlation difference at a given feature rank is $r_{\triangle} = r_{overlapping} - r_{non\text{-}overlapping}$. $r_{overlapping}$ is calculated by identifying all proteomic features at a given rank that have overlapping transcriptomic features, taking the correlation between each feature pair, and taking the average. Similarly, $r_{non\text{-}overlapping}$ is calculated by identifying all proteomic features at a given rank that do not have overlapping transcriptomic features, taking the correlation between each feature pair, and taking the average. This is normalized to the mean $r_{\triangle}$ across all feature ranks. Thus, a value of 1 represents the average $r_{\triangle}$ across all features after normalization (horizontal gray line). The top 500 features are demarcated by the red dashed line.

Next, we asked how the proteomic and transcriptomic data modalities were combined with regards to the model fit. During hyperparameter tuning, whereas all 19,138 features were selected for transcriptomics, only 500 of the 10,969 proteomic features were selected, representing 2.55% of all features used by the model. The selection of only a small subset of proteomic features but all transcriptomics features was consistent across model types (Table 3). However, proteomic features represent 3.40% of the top 500 features (odds ratio 1.36, Fisher's

exact test one-sided p-value 0.14) and are maximized at the 384th rank at 3.71% (odds ratio 1.49, Fisher's exact test one-sided p-value 0.10). of the feature set (Fig. 3b). While over-representation in the top 500 features is not significant ($p > 0.05$), it is trending towards significant (Fig. S8). These results indicate that while proteomics comprise only a small subset of the features, they represent important additional information for the model. This aligns with the potential boost in performance provided by using the joint omics data (Fig. 2b).

Finally, given the variance in correlation between mRNA and protein abundance[25,26], we asked the extent to which gene products overlapped between these two modalities. While there was nearly complete overlap across all features, only 35.3% of proteomic features overlapped with transcriptomic features in the top 500, indicating a significant depletion (odds ratio 0.02, Fisher's exact test one-sided p-value 7.77e-11) of gene product overlap between mRNA and protein (Fig. 3c). This indicates that the model may be accounting for redundant or complementary information between mRNA and protein. In instances where a proteomic feature was selected over its corresponding transcriptomic feature with a shared gene product at a given feature rank (non-overlapping), these may carry redundant information. This could indicate that the protein feature is more informative of cancer cell metastasis than the corresponding transcriptomic feature. In instances where both the proteomic feature and its corresponding transcriptomic feature with a shared gene product were included at a given feature rank (overlapping), there may be complementary information between the two modalities.[21] To explicitly test this, we looked at how the Pearson correlation between pairs of features differed between overlapping and non-overlapping features (Fig. 3d). We use higher Pearson correlation as a proxy for higher redundancy and lower Pearson correlation as a proxy for higher complementarity. While overlapping features almost always had higher Pearson correlation at a given rank (Fig. S9), the difference in Pearson correlation between overlapping and non-overlapping features tends to be lower than average at more important feature ranks. Similarly, it tends to be higher than average or close to average at more lowly ranked features (Fig. 3d). This supports the notion that, in highly ranked features, the model selects overlapping features when they are complementary and non-overlapping ones when they are redundant.

Discussion

In this study, we explored the extent to which transcriptomics and proteomics are informative of metastasis. Numerous studies have looked at the association between molecular profiles and cancer metastasis[14]. However, such approaches have not been used on the combination of transcriptomics and proteomics, to build interpretable linear models predicting cancer metastasis in a continuous manner, and in a generalizable manner that generalizes to multiple tumor types.

First, we observed that linear machine learning models can predict metastasis as well as more complex non-linear models, including ANNs. Given their higher interpretability, it is beneficial to identify features predictive of metastasis with linear models for metastatic studies of similar sample sizes. However, this speaks, in part, to the large variance and moderate predictive power of these omics data in metastasis, yielding what appears to be an upper limit to the

information standard machine learning approaches can extract from the data. Additional omics layers, such as DNA-methylation[16], and additional samples with more variance, which may be limited in cancer cell lines, may help increase predictive performance.

Next, we observed that transcriptomics predictions tend to outperform proteomic ones; this may in part be due to the measurement of much fewer genes relative to RNA-sequencing. Combining proteomics and transcriptomics did offer a slight boost in predictive performance. More interestingly, the model fits combined features from both modalities in a non-redundant manner: there was significant depletion in the overlap of gene products in the top 500 genes and those that did overlap had relatively lower correlation. This indicates that a subset of proteomic features is more informative than their redundant transcriptomic counterpart (those that don't overlap) and another subset offers information complementary to that available in transcriptomics (those that do overlap). This indicates that, across cell lines, the model fit is accounting for gene-specific variance in correlation[25] between mRNA and protein. Furthermore, the presence of non-linear relationships and network effects[30] between mRNA and protein can explain instances in which the two modalities are distinctly informative[21] of metastasis. Additionally, we note that the potential boost in predictive performance was small, indicating that RNA-sequencing alone, which is more widely accessible, will suffice for predicting metastatic phenotypes in similar datasets for most purposes.

Finally, we observed that the most informative features in our linear joint omics model were associated with relevant processes such as cell migration and chemotaxis. This indicates that, despite moderate predictive performance, the model is learning biologically meaningful connections between metastasis and the molecular profile of the cell lines.

These results are based on data generated from cancer cell lines, which might have substantial differences from *in vivo* tumors. *In vivo* metastasized tumors contain heterogeneous microenvironments influenced by mutational burden, interactions with tissue-specific cell types and the immune system, and nutrient availability and hypoxia[8,31]. Similarly, the metastatic potential was quantified in mice, and therefore might not directly translate to human metastatic phenotypes due to factors such as functional differences in orthologs[32], distinct global responses to perturbation[33], and pathogenesis through disease mechanisms that are subdominant in one species[34]. While molecular processes in these simpler models do not fully reflect *in vivo* responses, our pathway enrichment results indicate that the model still captures biologically relevant features.

Thus, exploring model structures that have both predictive power and can generalize across multiple contexts (e.g. species or experimental model types)[33] can improve interpretability and clinical relevance. Similarly, while our model is tissue-agnostic, explicitly accounting for the tissue-specific metastatic potential available in MetMap may improve specificity and predictive power. Finally, using *in silico* simulations to assess how perturbation of specific features alters metastatic potential can identify potential biomarkers or candidate targets for therapeutic intervention.

Methods

Transcriptomic and Proteomic Data Processing

Metastatic potential for 488 samples reported by MetMap 500[19] was downloaded from https://depmap.org/metmap/data/index.html.

Batch-corrected RNA-sequencing data of protein-coding genes for the DepMap[17] cell lines was downloaded in log-normalized transcripts per million (https://depmap.org/portal/data_page/?tab=allData&releasename=DepMap%20Public%2024Q4&filename=OmicsExpressionProteinCodingGenesTPMLogp1BatchCorrected.csv). 481 samples overlapped out of those reported in MetMap 500 and the 1673 reported in the RNA-sequencing dataset.

Normalized protein expression data as reported in Table S2 of Nusinow et al.[18] were used. Proteins with missing values in more than 80% of samples (1,786 of 12,755 features) were removed. The remaining missing values were imputed in Perseus (version 2.1.3.0) as follows: the protein distribution was transformed by 1) subtracting 1.8 standard deviations from all values and then 2) multiplying all values by 0.3 to shrink the standard deviation; next, imputed values were drawn by random sampling from this transformed distribution.

Three cell lines had two samples originating from different ten-plexes, which were dropped. Finally, 248 samples overlapped out of those reported in MetMap 500 and the 372 reported in the proteomics dataset. Finally, 247 samples were present in both the transcriptomics and proteomics data.

Model Fitting and Hyperparameter Tuning

Models were either fit on transcriptomics separately, proteomics separately, or the two datasets jointly. The following preprocessing steps were always conducted on either modality separately: the top features were selected according to those that had the highest coefficient of variation (training data only in cases of train-test splits), proceeded by mean-centering. The number of features used by the model was one of the hyperparameters explored (see Table 3 for a comprehensive accounting of hyperparameters explored).

The following linear regression models were each fit on the data, with omics as the input and metastatic potential as the output: partial least squared (PLS) regression, Ridge, Lasso, ElasticNet, Support Vector Machine (SVM) with a linear kernel. Similarly, the following nonlinear regression models were fit on the data: SVM with polynomial and radial basis function kernels, k-Nearest Neighbors (KNN), and Random Forest.

Each model type was fit on the same 10 train-test splits using 10-fold CV. Within each fold, an inner 5-fold CV on the training data was used to identify the best set of hyperparameters (Table

3, sheet 1). Hyperparameter tuning was conducted using Optuna[20] with 100 trials to minimize the average mean squared error. The “SuccessiveHalvingPruner” pruner was used alongside a custom sampler. The custom sampler used the CmaEsSampler on hyperparameters when possible, and the TPESampler otherwise. To prevent convergence to a local minimum, every 20 trials, hyperparameters were also randomly sampled using the TPESampler.

Upon identification of the best set of hyperparameters, model predictions were assessed on the outer fold test data using Pearson correlation and mean squared error (Fig. 1a, Fig. S3). This resulted in a unique set of hyperparameters for each fold for each model type (Table 3, sheets 2-4). A single consensus model for each of the top three performing model types (by mean across folds) for both performance metrics was generated as follows: 1) for discrete and integer hyperparameters, the mode across 10-folds was used, and 2) for continuous hyperparameters, the mean value across the 10-folds was used. The exception here was the number of features selected for proteins in the joint omics pipeline, in which the median rather than the mode was used because there were more uniform distributions across folds. Once a consensus model was identified, it was assessed using the same performance metrics on a different set of folds using 10-fold CV.

Neural Network Construction and Training

For transcriptomics, alongside the consensus models, we also fit single and ensemble artificial feed-forward neural networks (ANNs) on the 10-fold CV. Single ANNs consist of one initial dropout layer, with a dropout rate of 0.5, and four blocks which contain in series: a fully-connected linear layer, a batch normalization layer with (momentum=0.2), a non-linear activation function (ELU), and a dropout layer with a dropout rate of 0.25. The final layer is a Gaussian regression layer that predicts the mean ($\bar{y}_i$) and a variance (y_{σ_i}) of the metastatic potential, rather than a single value. This treats the metastatic potential as a sample from a Gaussian distribution with the predicted mean and variance, and the model is trained by minimizing the negative log-likelihood criterion[35] given by: $-\ log(p_\theta|X_n) = -\frac{1}{2}log\sigma_\theta^{\ 2}(x) - \frac{1}{2\sigma_\theta^{\ 2}(x)}(y - \mu_\theta(x))^2 + constant$ and minimized through the following loss function: $Loss = \frac{1}{2}logy_{\sigma_i} + \frac{1}{2y_{\sigma_i}}(y_{true} - \bar{y}_i)^2 + 10^{-6}$. We note that for the power analysis (see next section for details), we do not use a Gaussian regression layer as the final layer, but rather a simple linear feed-forward neural network layer. Each layer reduces the dimensionality of the data progressively to 1024, 256, 32, and 16 nodes. Models were trained for 100 epochs with a learning rate of 0.001 and a batch size of 40 using the Adam optimizer. To avoid overfitting L2-regularization was applied to the models’ parameters with a regularization strength of 0.01.

Next, an ensemble of 20 of these ANNs is trained. The output of this ensemble is also a Gaussian, with mean and variance calculated from the uniformly weighted mixture of each

model. The final prediction of the model corresponds to the mean of the mixture: $y_{\mu} = \frac{1}{N}\sum_{i=1}^{N} \bar{y}_i$.

The variance of the mixture is defined as: $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_{\sigma_i} + \bar{y}_i^{\,2}) - y_{\mu}^{\,2}}$.

Power Analysis of Transcriptomic Model

To assess how the top-performing, transcriptomics-fit consensus models are affected by sample size, we conducted a power analysis. We first split data into training and test splits using 10-fold CV. For each fold, we subsetted training data between 10% to 100% of the sample size at intervals of 10%. We also conducted the power analysis at the sample sizes of the proteomics and joint omics datasets (248 and 247 samples respectively). With the exception of the entire training sample size, for each interval, we selected 100 random subsets to fit the model on, resulting in a total of 1000 test predictions per training sample size.

Next, we quantified the relationship between prediction performance, sample size, and model type. First, we determined whether a linear, power, or exponential function best explains the relationship between model performance and sample size. To do so, we fit linear models for each of MSE or Pearson correlation regressed on sample size for each model type individually. Prior to model fitting, for Pearson correlation, we applied a Fisher's z-transformation. For power functions, we fit the relationship $m = aN^b$, where m is the model performance (Fisher z-transformed Pearson correlation or MSE), N is the training sample size, and a and b are the parameters for the power function. We log-transform this data to fit a linear regression to it as follows: $\log_{10}(m) = \log_{10}(a) + b*\log(N)$. Similarly, for exponential functions, we fit the relationship $m = a*e^{bN}$. We log-transform this data to fit a linear model to it as follows: $\log_e(m) = \log_e(a) + bN$. Next, we assessed goodness of fit using the adjusted R^2, Akaike information criterion (AIC), and Bayesian information criterion (BIC). We observed that on average across the predictive models, power functions best explained performance by adjusted R^2 for both Pearson correlation and MSE, and they also best explained performance by AIC and BIC for MSE (Table 2).

As such, we proceeded with a power function to fit the following regression in order to assess differences in model types' predictive performance across sample size: $\log_{10}(m) = \log_{10}(a) + b*\log(N) + model_i$, where i is the model type and the Linear SVR is the reference model. When back-transforming to the power function, these fitted parameters can be interpreted as follows: $m = aN^{b*}10^{model_i}$. For the Linear SVR to be the reference, $model_linearSVR = 0$, such that $m = aN^{b*}10^0$, which simplifies to $m = aN^b$. Thus, each other model covariate is a scaling of model performance by 10^{model_i} relative to this reference.

References

1. Guan, X. Cancer metastases: challenges and opportunities. *Acta Pharm Sin B* **5**, 402–418

(2015).
2. Chaffer, C. L. & Weinberg, R. A. A perspective on cancer cell metastasis. *Science* **331**, 1559–1564 (2011).
3. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
4. D'Orazio, M. *et al.* Deciphering cancer cell behavior from motility and shape features: Peer prediction and dynamic selection to support cancer diagnosis and therapy. *Front. Oncol.* **10**, 580698 (2020).
5. Stuelten, C. H., Parent, C. A. & Montell, D. J. Cell motility in cancer invasion and metastasis: insights from simple model organisms. *Nat. Rev. Cancer* **18**, 296–312 (2018).
6. Shi, X. *et al.* Mechanism insights and therapeutic intervention of tumor metastasis: latest developments and perspectives. *Signal Transduct. Target. Ther.* **9**, 192 (2024).
7. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* **34**, 1145–1160 (2016).
8. Baghdassarian, H. M. & Lewis, N. E. Resource allocation in mammalian systems. *Biotechnol Adv* **71**, 108305 (2024).
9. Lewis, N. E., Nagarajan, H. & Palsson, B. O. Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat Rev Microbiol* **10**, 291–305 (2012).
10. Chuang, H.-Y., Hofree, M. & Ideker, T. A decade of systems biology. *Annu Rev Cell Dev Biol* **26**, 721–744 (2010).
11. Subramanian, A. *et al.* A next generation Connectivity Map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452.e17 (2017).
12. Reel, P. S., Reel, S., Pearson, E., Trucco, E. & Jefferson, E. Using machine learning approaches for multi-omics data analysis: A review. *Biotechnol Adv* **49**, 107739 (2021).
13. Zhang, Z. *et al.* Morphology-based prediction of cancer cell migration using an artificial neural network and a random decision forest. *Integr. Biol.* **10**, 758–767 (2018).
14. Albaradei, S. *et al.* Machine learning and deep learning methods that use omics data for metastasis prediction. *Comput. Struct. Biotechnol. J.* **19**, 5008–5018 (2021).
15. Lee, S. C., Quinn, A., Nguyen, T., Venkatesh, S. & Quinn, T. P. A cross-cancer metastasis signature in the microRNA-mRNA axis of paired tissue samples. *Mol. Biol. Rep.* **46**, 5919–5930 (2019).
16. Albaradei, S. *et al.* MetaCancer: A deep learning-based pan-cancer metastasis prediction model developed using multi-omics data. *Comput. Struct. Biotechnol. J.* **19**, 4404–4411 (2021).
17. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564–576.e16 (2017).
18. Nusinow, D. P. *et al.* Quantitative Proteomics of the Cancer Cell Line Encyclopedia. *Cell* **180**, 387–402.e16 (2020).
19. Jin, X. *et al.* A metastasis map of human cancer cell lines. *Nature* **588**, 331–336 (2020).
20. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. (2019).
21. Buccitelli, C. & Selbach, M. mRNAs, proteins and the emerging principles of gene expression control. *Nat Rev Genet* **21**, 630–644 (2020).
22. Yamaguchi, H., Wyckoff, J. & Condeelis, J. Cell migration in tumors. *Curr. Opin. Cell Biol.* **17**, 559–564 (2005).
23. Hood, J. D. & Cheresh, D. A. Role of integrins in cell invasion and migration. *Nat. Rev. Cancer* **2**, 91–100 (2002).
24. Tochhawng, L., Deng, S., Pervaiz, S. & Yap, C. T. Redox regulation of cancer cell migration and invasion. *Mitochondrion* **13**, 246–253 (2013).
25. Edfors, F. *et al.* Gene-specific correlation of RNA and protein levels in human cells and tissues. *Mol Syst Biol* **12**, 883 (2016).

26. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587 (2014).
27. Alwosheel, A., van Cranenburgh, S. & Chorus, C. G. Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. *J. Choice Model.* **28**, 167–182 (2018).
28. Belkin, M., Hsu, D., Ma, S. & Mandal, S. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 15849–15854 (2019).
29. M., C. & Hugh. *Deep Learning: Foundations and Concepts*. (Springer International Publishing, Cham, Switzerland, 2023).
30. Sweatt, A. J., Griffiths, C. D., Paudel, B. B. & Janes, K. A. Proteome-wide copy-number estimation from transcriptomics. *bioRxiv* (2023) doi:10.1101/2023.07.10.548432.
31. Binnewies, M. *et al.* Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nat Med* **24**, 541–550 (2018).
32. Liao, B.-Y. & Zhang, J. Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proc Natl Acad Sci U S A* **105**, 6987–6992 (2008).
33. Meimetis, N. *et al.* AutoTransOP: translating omics signatures without orthologue requirements using deep learning. *NPJ Syst. Biol. Appl.* **10**, 13 (2024).
34. Pullen, K. M. *et al.* Cross species modeling reveals a role for the unfolded protein response in shaping the transcriptional reaction to Mycobacterium tuberculosis infection. *bioRxiv* 2024.04.19.590290 (2024) doi:10.1101/2024.04.19.590290.
35. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. in *Advances in Neural Information Processing Systems* 6402–6413.

Acknowledgements

We would like to thank Luka Karginov and Brian Joughin for helpful discussions regarding partial correlation networks and feature selection, respectively. H.M.B is funded as a MIT-Novo Nordisk Artificial Intelligence Postdoctoral Fellow. N.M. is a 2024-2025 Takeda Fellow. The funders played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Ethics Declaration

Competing Interests

The authors declare no competing interests.

Data Availability

All data is publicly available through DepMap (https://depmap.org/portal/) and CCLE (https://sites.broadinstitute.org/ccle/).

Supplementary Figures

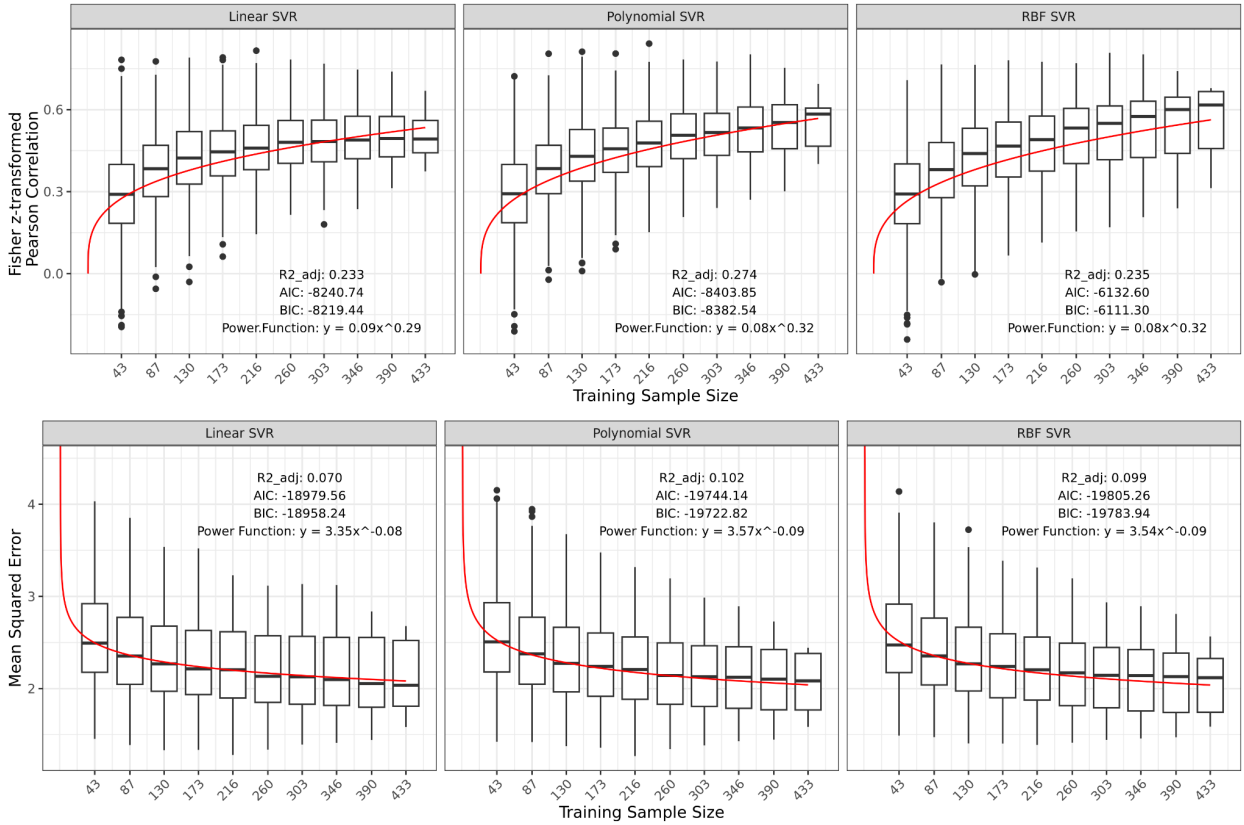

Fig. S1: Results of power analysis on consensus transcriptomic models. Box Plots of individual consensus models fit on transcriptomics data. Model performance (y-axis) assessed by Fisher z-transformed Pearson correlation (top panels) or mean squared error (bottom panels) across training sample size (x-axis). Red lines visualize the power function regression fits for each model type and performance metric. Plots are annotated with assessment metrics for these regression fits as well as regression parameters.

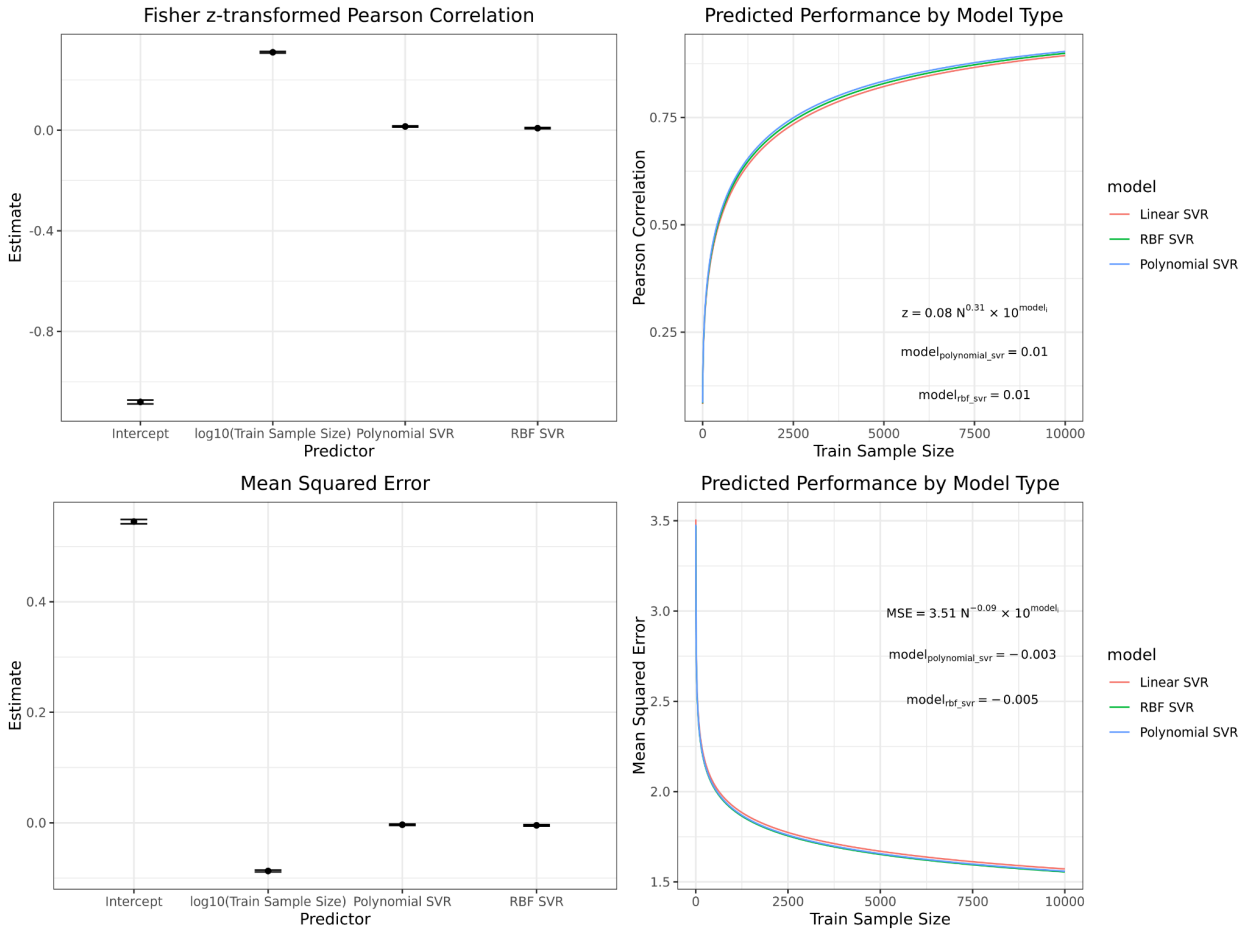

Fig. S2: Power function fits for performance of consensus transcriptomic models across sample size. Performance of consensus models by Pearson correlation (top panels) and MSE (bottom panels) estimated by power function regression fits that include the model type as a covariate. Left panels show estimates (y-axis, scatter point) and standard errors (horizontal bars) for each parameter (x-axis) in the model prior to back transformation (exponentiation and, in the case of Pearson correlation, inverse Fisher z-transformation). All estimates are significant (two-sided Student's t-test, p-value ≤ 0.004). In a power function $y = ax^b$, the intercept represents $\log_{10}(a)$, "log10(Train Sample Size)" represents b, and the remaining estimates represent the effect of the model type with respect to the Linear SVR as a baseline (see Methods for details). Right panels visualize predicted performance (y-axis) for each model type across training sample size (x-axis) from 0 to 10,000 training samples. Plots are annotated with back-transformed power function fitted parameters, with N representing training sample size and model_i representing the effect of the model type with respect to Linear SVR ($\text{model}_{\text{linear_svr}} = 0$, $10^0 = 1$, no effect). The correlation function annotation is in terms of the Fisher z-transformed Pearson correlation (z), whereas the curves and y-axis are back-transformed to Pearson correlation.

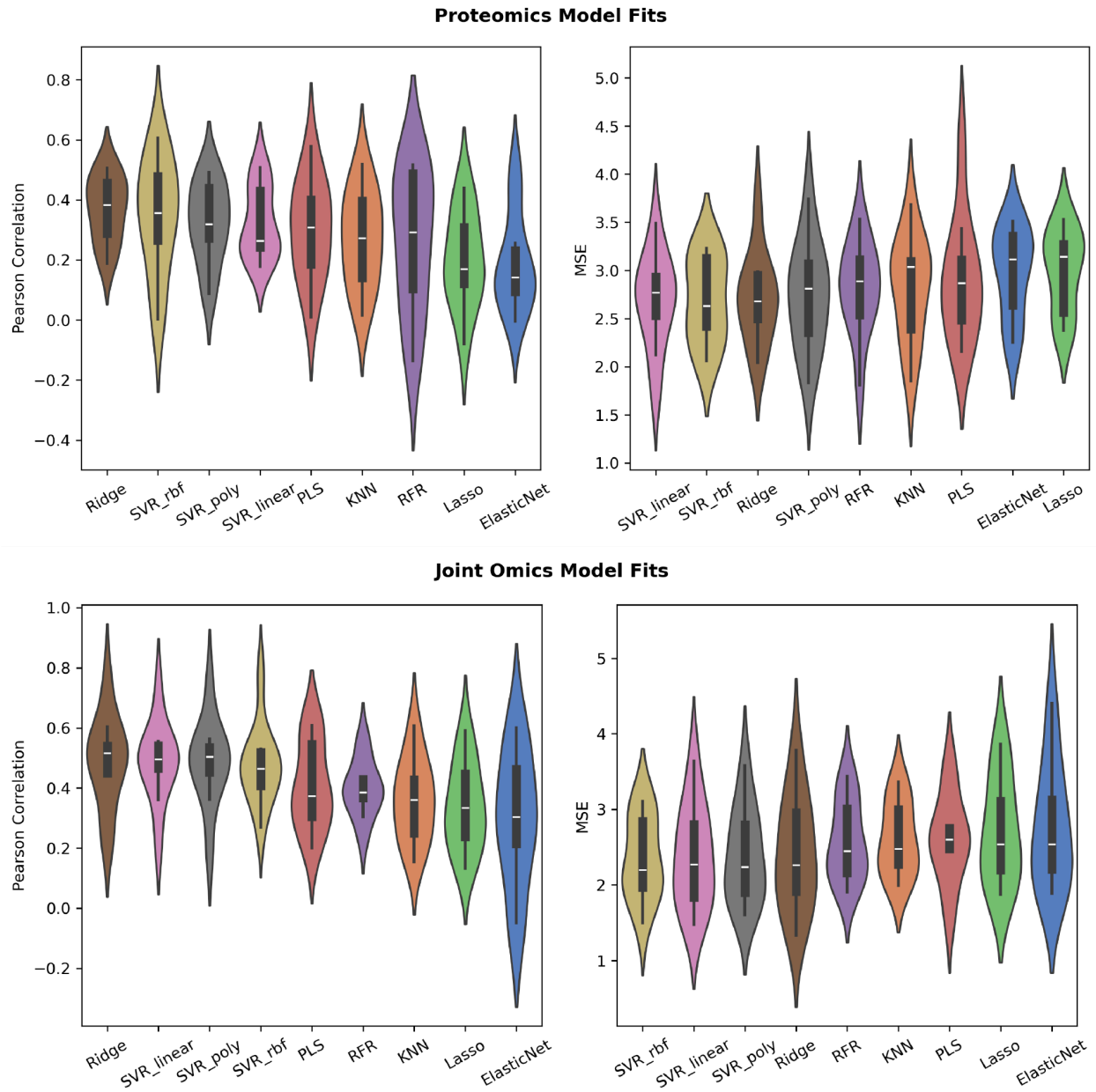

Fig. S3: Results of 10-fold CV for all hyperparameter tuned models fit on proteomics data. All statistical significance is determined using a two-sided Mann-Whitney U (MWU) test and the FDR is controlled for multiple testing using the Benjamini-Hochberg correction (**** $q \leq 10^{-4}$, *** $q \leq 10^{-3}$, ** $q \leq 0.01$, * $q \leq 0.1$). Comparisons that are not significant are not annotated. Models' predictive performance is assessed either by Pearson correlation (left panel) or mean squared error (right panel). Violin plots visualize performance distribution output from 10-fold CV. Performance of 9 predictive models (x-axis) is ordered by best mean value. For each fold, specific model hyperparameters were optimized using nested 5-fold CV. No comparisons have FDR ≤ 0.1. This is analogous to Fig. 1a in the main text for **a** proteomics and **b** transcriptomic and proteomics combined.

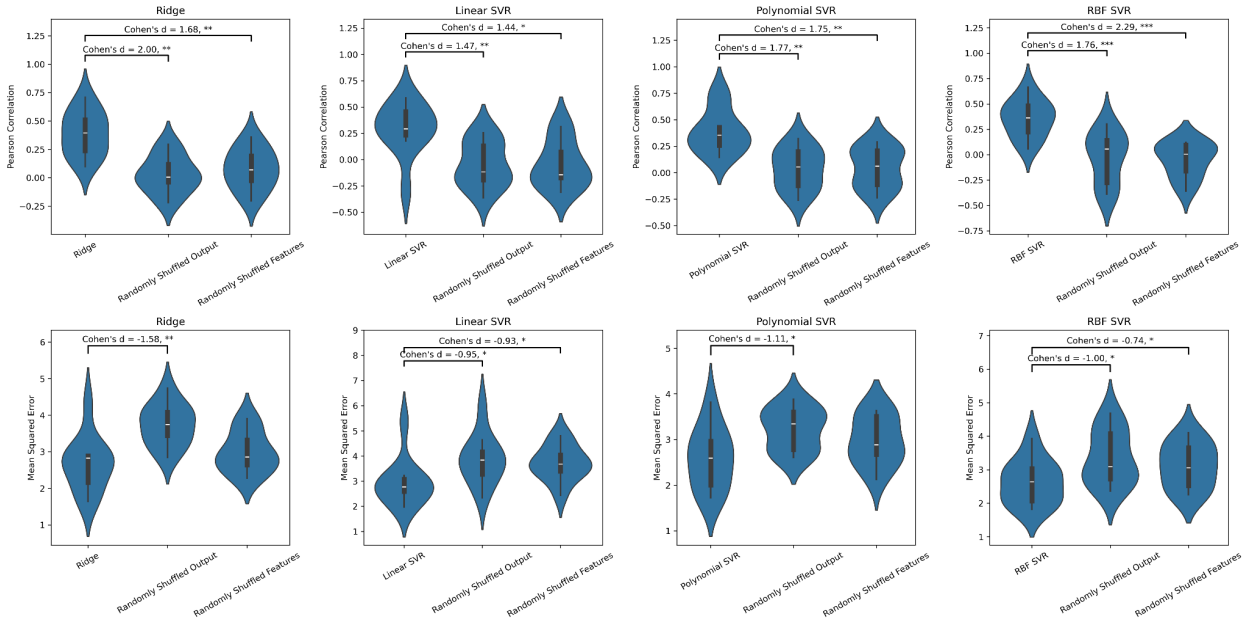

Fig. S4: Comparison of consensus proteomics models to random baselines. All statistical significance is determined using a Mann-Whitney U (MWU) test and the FDR is controlled for multiple testing using the Benjamini-Hochberg correction (**** $q \leq 10^{-4}$, *** $q \leq 10^{-3}$, ** $q \leq 0.01$, * $q \leq 0.1$). Comparisons that are not significant are not annotated. Statistical tests are one-sided (testing whether the actual model is less than random) unless otherwise specified. Models' predictive performance is assessed either by Pearson correlation (top panels) or mean squared error (bottom panels). Violin Plots display performance distributions output from 10-fold CV.

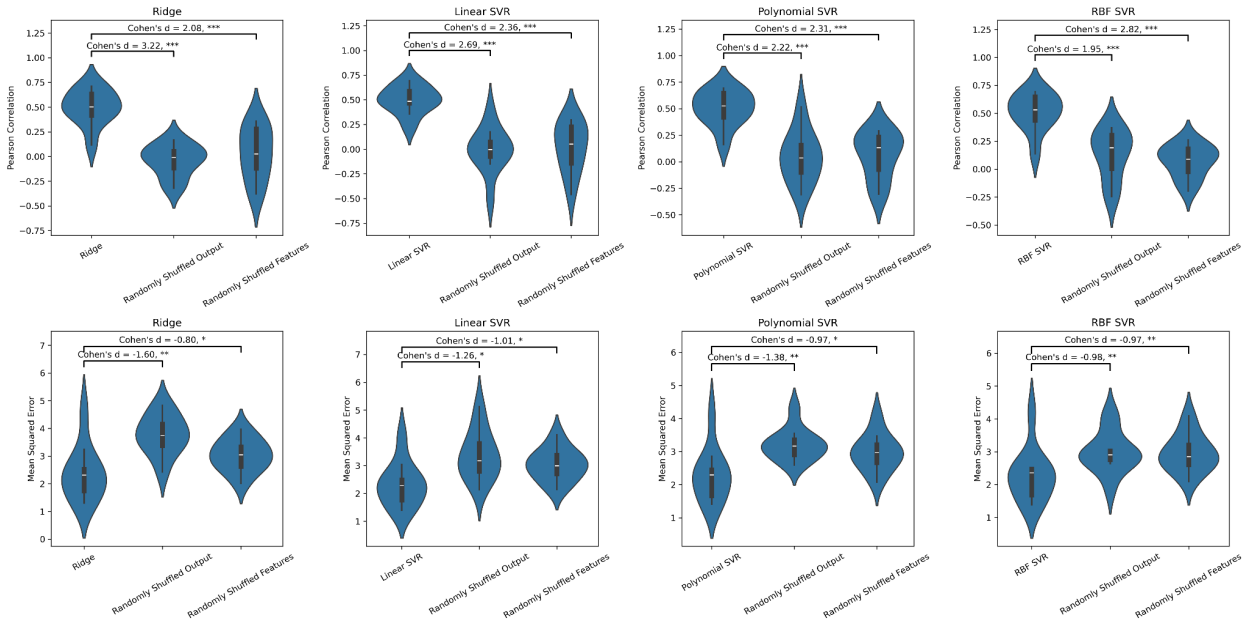

Fig. S5: Comparison of consensus joint omics models to random baselines. All statistical significance is determined using a Mann-Whitney U (MWU) test and the FDR is controlled for multiple

testing using the Benjamini-Hochberg correction (**** $q \leq 10^{-4}$, *** $q \leq 10^{-3}$, ** $q \leq 0.01$, * $q \leq 0.1$). Comparisons that are not significant are not annotated. Statistical tests are one-sided (testing whether the actual model is less than random). Models' predictive performance is assessed either by Pearson correlation (top panels) or mean squared error (bottom panels). Violin Plots display performance distributions output from 10-fold CV.

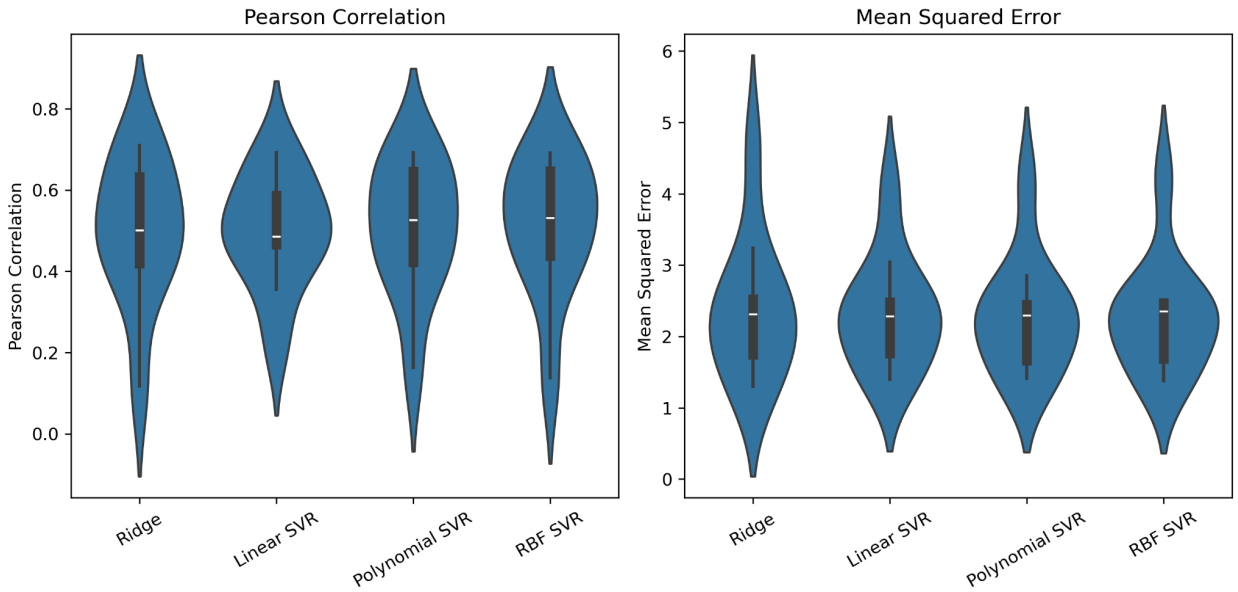

Fig. S6: Comparison of performance between consensus joint omics models. All statistical significance is determined using a Mann-Whitney U (MWU) test and the FDR is controlled for multiple testing using the Benjamini-Hochberg correction (**** $q \leq 10^{-4}$, *** $q \leq 10^{-3}$, ** $q \leq 0.01$, * $q \leq 0.1$). Comparisons that are not significant are not annotated. Statistical tests are two-sided. Models' predictive performance is assessed either by Pearson correlation (left panel) or mean squared error (right panel). Violin Plots display performance distributions output from 10-fold CV.

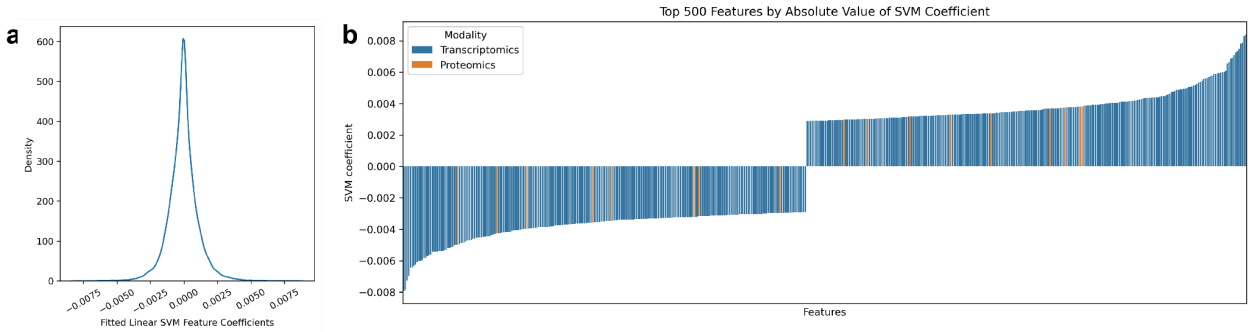

Fig. S7 Top 500 joint omics consensus linear SVR features. a Kernel density estimate of joint omics linear SVR coefficients for all features. **b** Barplot of top 500 model coefficients (y-axis) rank-ordered by coefficient value (x-axis) and colored by data modality.

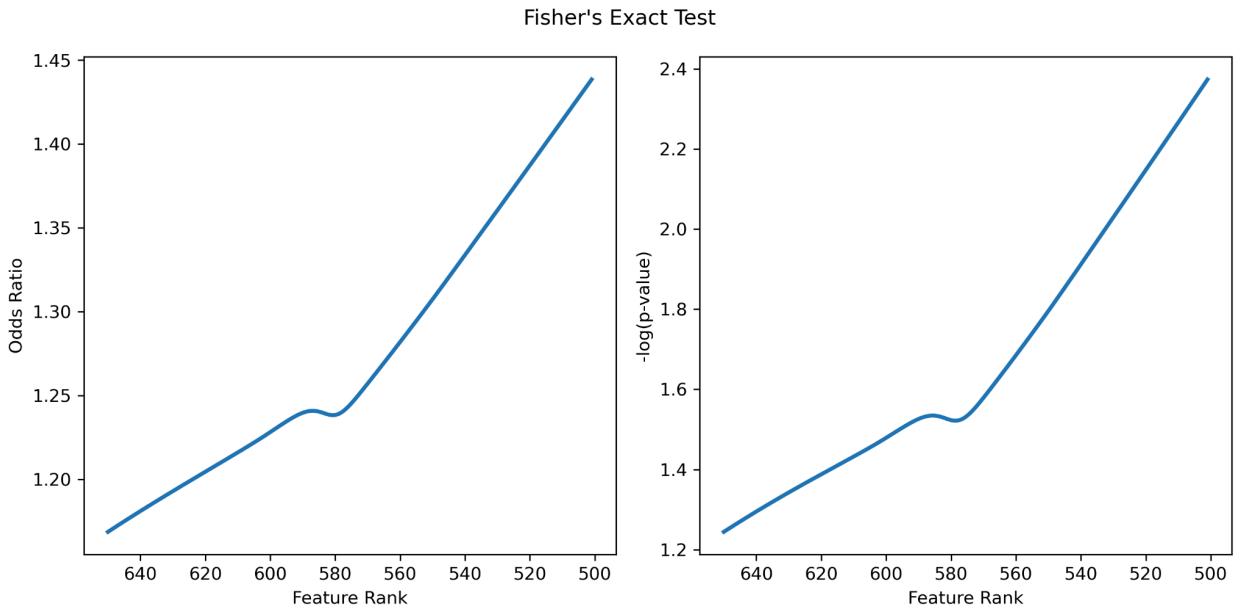

Fig. S8 Trends for over-representation of proteomic features in those that are top-ranked. Odds ratio (y-axis, left panel) and -log(p-value) from one-sided Fisher's exact test (y-axis, right panel) testing for over-representation of proteomic features at each feature rank from the top 650 features to the top 500 features (x-axis). Curves are non-parametric local weighted linear regressions (lowess) fit on the feature rank.

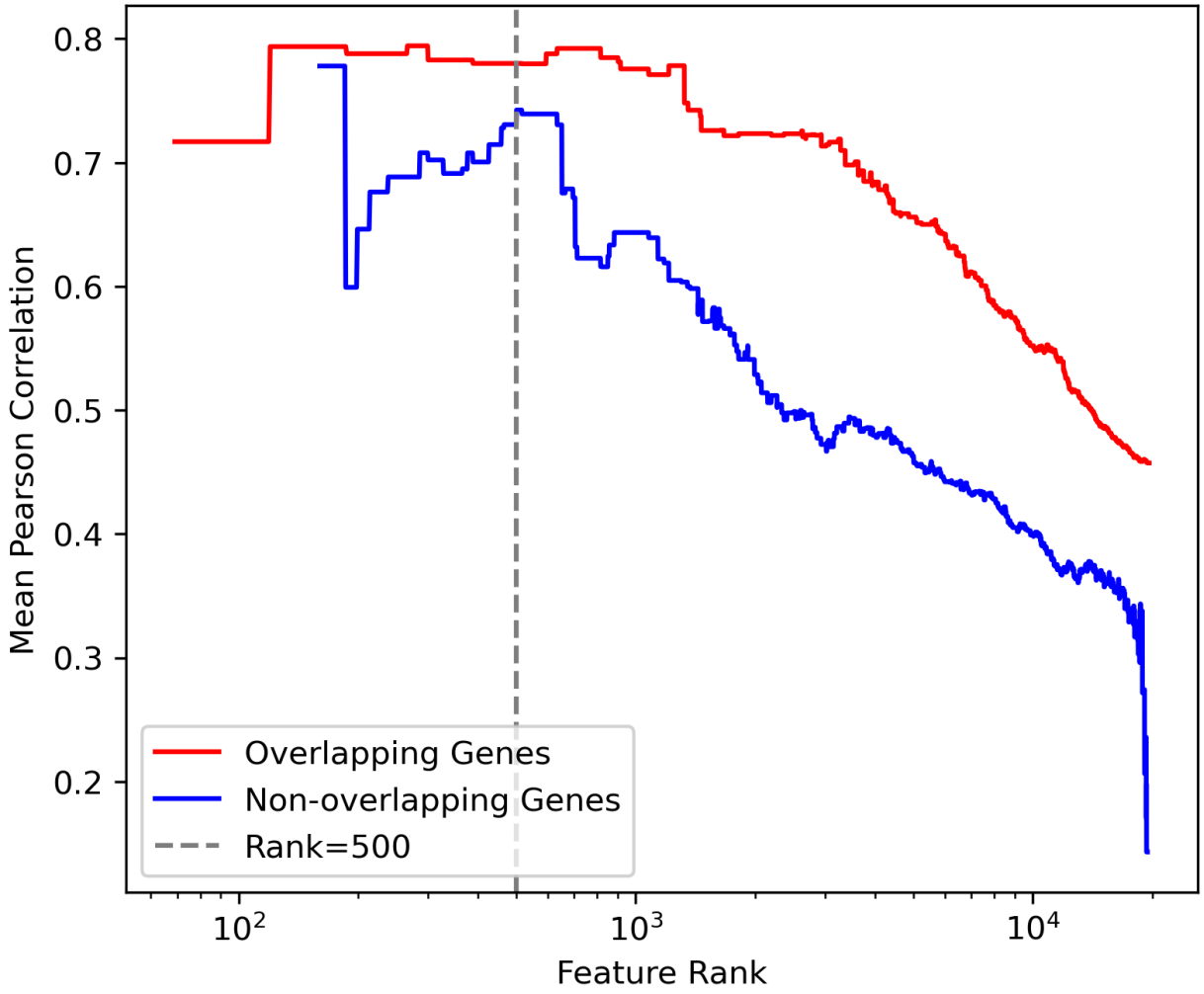

Fig. S9 Comparison of protein-mRNA feature correlations at a given feature rank. Line Plots comparing proteomic-to-transcriptomic feature correlation for overlapping or non-overlapping genes (y-axis) at a given rank-order (x-axis). The red and blue lines represent $r_{overlapping}$ and $r_{non\text{-}overlapping}$, respectively, as defined in Fig. 3d